

THE MOST PARSIMONIOUS TREE FOR RANDOM DATA

MAREIKE FISCHER, MICHELLE GALLA, LINA HERBST AND MIKE STEEL

ABSTRACT. Applying a method to reconstruct a phylogenetic tree from random data provides a way to detect whether that method has an inherent bias towards certain tree ‘shapes’. For maximum parsimony, applied to a sequence of random 2-state data, each possible binary phylogenetic tree has exactly the same distribution for its parsimony score. Despite this pleasing and slightly surprising symmetry, some binary phylogenetic trees are more likely than others to be a most parsimonious (MP) tree for a sequence of k such characters, as we show. For $k = 2$, and unrooted binary trees on six taxa, any tree with a caterpillar shape has a higher chance of being an MP tree than any tree with a symmetric shape. On the other hand, if we take any two binary trees, on any number of taxa, we prove that this bias between the two trees vanishes as the number of characters grows. However, again there is a twist: MP trees on six taxa are more likely to have certain shapes than a uniform distribution on binary phylogenetic trees predicts, and this difference does not appear to dissipate as k grows.

Keywords: Tree, maximum parsimony, random data, central limit theorem

1. INTRODUCTION

The ‘shape’ of reconstructed evolutionary trees is of interest to evolutionary biologists, as it should provide some insight into the processes of speciation and extinction [2, 3, 10, 11, 12, 14]. In this paper, ‘shape’ refers just to the discrete shape of the tree (i.e. we ignore the branch lengths); the advantages of this are that it simplifies the analysis, and it also confers a certain robustness (i.e. the resulting probability distribution on discrete shapes is often independent of the fine details of an underlying speciation/extinction model [1], [12]). For example, if all speciation (and extinction) events affect all taxa at any given epoch in the

Date: June 3, 2014.

same way, then we should expect the shape of a reconstructed tree to be that predicted by the discrete ‘Yule–Harding’ model [2, 8, 12]. In fact, a general trend (see e.g. [2]) is that the shape of phylogenetic trees reconstructed from biological data tends to be a little less balanced than this model predicts, but is more balanced than what would be obtained under a uniform model in which each binary phylogenetic tree has the same probability (this model is sometimes also called the ‘Proportional-to-Distinguishable-Arrangements’ (PDA) model).

There are, however, other factors which can lead to biases in tree shape. One is non-random sampling of the taxa on which to construct a tree (influenced, for example, by the particular interests of the biologists or the application of a certain strategy to sample taxa). Another cause of possible bias is that a tree reconstruction method may itself have an inherent preference towards certain tree shapes. A way to test this latter possibility is to apply the tree reconstruction method to data that contain no phylogenetic signal at all, in particular, purely random data, where each character is generated independently by a process that assigns states to the taxa uniformly (e.g. by the toss of a fair coin in the case of two states). For some methods, such as ‘TreePuzzle’, such data leads to very balanced trees (similar to the Yule–Harding model [16, 17]). However, other methods, such as maximum likelihood and maximum parsimony, lead to less balanced trees, that are closer in shape to the uniform model, as recently reported in [11]. In the case of maximum parsimony, the two-state symmetric model has the even-handed property that every binary tree has exactly the same distribution of its parsimony score on k randomly generated characters. Thus, it might be supposed that the maximum parsimony (MP) tree for such a sequence of characters would also follow a uniform distribution. However, while this holds in special cases, it does not hold in general, as we show below.

1.1. Trees and parsimony: definitions and basic properties. In phylogenetics, graphs, especially trees, are used to describe the ancestral relationships among different species. A main goal of phylogenetics is to infer an evolutionary tree from data available from present-day species. In graph theory, a tree $T = (V, E)$ consists of a connected graph with no cycles.

Certain leaf-labelled trees (‘phylogenetic trees’) are widely used where the set of extant species label the leaves and the remaining vertices represent ancestral speciation events [5]. There are different methods of reconstructing a phylogenetic tree. One of the most famous tree reconstruction methods is maximum parsimony. For a given tree and discrete character data, the parsimony score can be found in polynomial time by using the Fitch–Hartigan algorithm [6, 9]. The parsimony score counts the number of changes (mutations) required on the tree to describe the data. This problem of finding the optimal parsimony score for a given tree is often called the ‘small parsimony’ problem. The ‘big parsimony’ problem aims at finding the most parsimonious tree (‘MP tree’) amongst all possible trees. This problem has been proven to be NP-hard [7].

In this paper, we assume that each taxon from the leaf set X of the tree is assigned a binary state (0 or 1) independently, and with equal probability. This process is then repeated (also independently) to generate a sequence of characters (defined formally below). For binary trees with random data, we are interested in the probability that a tree is an MP tree, and also what happens when the length of the sequences or the number of leaves gets larger. In particular, we wish to determine whether each tree is equally likely to be selected as an MP tree.

Definition 1. [Binary phylogenetic trees] An (*unrooted*) *binary phylogenetic X -tree* is a tree T with leaf set X and with every interior (i.e. non-leaf) vertex of degree exactly three. We will let $UB(X)$ be the set of unrooted binary phylogenetic X -trees. When $X = [n] = \{1, \dots, n\}$, we will write $UB(n)$.

Definition 2. [Character, extension, parsimony score]

- A *character on X* over a finite set R of character states is any function f from X into R ; $f : X \rightarrow R$. In this paper we will consider two-state characters; $f : X \rightarrow \{0, 1\}$.
- A function $\bar{f} : V \rightarrow R$ such that $\bar{f}|_X = f$ is said to be an *extension* of f since it describes an assignment of states to all vertices of T that agrees with the states that f stipulates at the leaves.

- Let $ch(\bar{f}, T) := |\{e = \{u, v\} \in E : \bar{f}(u) \neq \bar{f}(v)\}|$ be the *changing number* of \bar{f} . Given a character $f : X \rightarrow R$, the *parsimony score* of f on T , denoted $ps(f, T)$, is the smallest changing number of any extension of f , i.e. :

$$ps(f, T) := \min_{\bar{f}: V \rightarrow R, \bar{f}|_X = f} \{ch(\bar{f}, T)\}.$$

An extension \bar{f} of f for which $ch(\bar{f}, T) = ps(f, T)$ is said to be a *minimal extension*.

Let $\mathcal{C} = (f_1, \dots, f_k)$ be a sequence of characters on X . The *parsimony score* of \mathcal{C} on T , denoted $ps(\mathcal{C}, T)$, is defined by $ps(\mathcal{C}, T) := \sum_{i=1}^k ps(f_i, T)$.

2. COMPARING GIVEN TREES

Let $X_k(T)$ be the parsimony score of k random two-state characters on $T \in UB(n)$. We will see shortly (Proposition 1) that the distribution of $X_k(T)$ does not depend on the shape of T ; it just depends on n . Notice that $X_k(T) = X_1 + X_2 + \dots + X_k$, where X_i (for $i = 1, \dots, k$) form a sequence of independent and identically distributed random variables (with common distribution $X_1(T)$). If $\mathbb{P}(X_k(T) = l)$ denotes the probability that T has parsimony score l then, from [15], we have, for each $l \in [1, \lfloor n/2 \rfloor]$:

$$(1) \quad \mathbb{P}(X_1(T) = l) = \frac{2n - 3l}{l} \cdot \binom{n - l - 1}{l - 1} \cdot 2^{l-n},$$

with $\mathbb{P}(X_1(T) = 0) = 2^{1-n}$ and $\mathbb{P}(X_1(T) = l) = 0$ for $l > \lfloor n/2 \rfloor$. Furthermore, $\mathbb{E}[X_1(T)] = \frac{3n-2-(-\frac{1}{2})^{n-1}}{9} \sim \frac{n}{3}$ is the expected parsimony score of T , and $\mathbb{E}[X_k(T)] = k \cdot \mathbb{E}[X_1(T)]$. An immediate consequence of (1) is the following.

Proposition 1. For every $k \geq 1$ and $n \geq 2$, the distribution of the parsimony score of k independent random binary characters (i.e. $X_k(T)$) is the same for all $T \in UB(n)$.

2.1. Comparing two trees by their parsimony score. We begin this section by describing a tree rearrangement operation on binary phylogenetic trees [13, Chapter 2.6], namely tree bisection and reconnection (TBR). Let T be a binary phylogenetic X -tree and

let $e = \{u, v\}$ be an edge of T . A TBR operation is described as follows. Let T' be the binary tree obtained from T by deleting e , adding an edge between a vertex that subdivides an edge of one component of $T \setminus e$ and a vertex that subdivides an edge of the other component of $T \setminus e$, and then suppressing any resulting degree-two vertices. In the case that a component of $T \setminus e$ consists of a single vertex, then the added edge is attached to this vertex. T' is said to be obtained from T by a single TBR operation.

Proposition 2. Let $T, T' \in UB(n)$.

- (i) If T and T' are one TBR apart, then $\mathbb{P}(X_k(T) < X_k(T')) = \mathbb{P}(X_k(T') < X_k(T))$ holds for all $k \geq 1$.
- (ii) If T and T' are more than one TBR apart, then the equality $\mathbb{P}(X_k(T) < X_k(T')) = \mathbb{P}(X_k(T') < X_k(T))$ can fail, even for $k = 1$ and $n = 6$.

Proof.

- (i) From [4, Lemma 5.1], if T and T' are one TBR apart then for any character f , $|ps(f, T) - ps(f, T')| \leq 1$. In particular,

$$(2) \quad |X_1(T) - X_1(T')| \leq 1.$$

For $k \geq 1$, let $\Delta_k = X_k(T) - X_k(T')$. Then if $T, T' \in UB(n)$ are one TBR apart, then $\Delta_1 = X_1(T) - X_1(T')$ is either 0, 1 or -1 , by (2). Moreover, $\mathbb{P}(\Delta_1 = m) = \mathbb{P}(\Delta_1 = -m)$ for all $m \in \{0, 1 - 1\}$, since $\mathbb{E}[\Delta_1] = 0$, by Proposition 1. Furthermore, $\Delta_k = D_1 + \cdots + D_k$, where D_1, \dots, D_k are independent and identically distributed

as Δ_1 , so we have:

$$\begin{aligned}
\mathbb{P}(\Delta_k = m) &= \sum_{\substack{m_1, \dots, m_k \in \{-1, 0, 1\}: \\ m_1 + \dots + m_k = m}} \mathbb{P}(D_1 = m_1 \wedge D_2 = m_2 \wedge \dots \wedge D_k = m_k) \\
&= \sum_{\substack{m_1, \dots, m_k \in \{-1, 0, 1\}: \\ m_1 + \dots + m_k = m}} \prod_{j=1}^k \mathbb{P}(D_j = m_j) = \sum_{\substack{m_1, \dots, m_k \in \{-1, 0, 1\}: \\ m_1 + \dots + m_k = m}} \prod_{j=1}^k \mathbb{P}(D_j = -m_j) \\
&= \sum_{\substack{m'_1, \dots, m'_k \in \{-1, 0, 1\}: \\ m'_1 + \dots + m'_k = -m}} \mathbb{P}(D_1 = m'_1 \wedge D_2 = m'_2 \wedge \dots \wedge D_k = m'_k) = \mathbb{P}(\Delta_k = -m).
\end{aligned}$$

This provides the equality $\mathbb{P}(X_k(T) < X_k(T')) = \mathbb{P}(X_k(T') < X_k(T))$ for all $k \geq 1$.

(ii) We prove this by exhibiting one counterexample, namely the trees shown in Fig. 1.

Let $\Delta_k = X_k(T) - X_k(T')$. The equality $\mathbb{P}(X_k(T) < X_k(T')) = \mathbb{P}(X_k(T') < X_k(T))$ is equivalent to $\mathbb{P}(\Delta_k < 0) = \mathbb{P}(\Delta_k > 0)$.

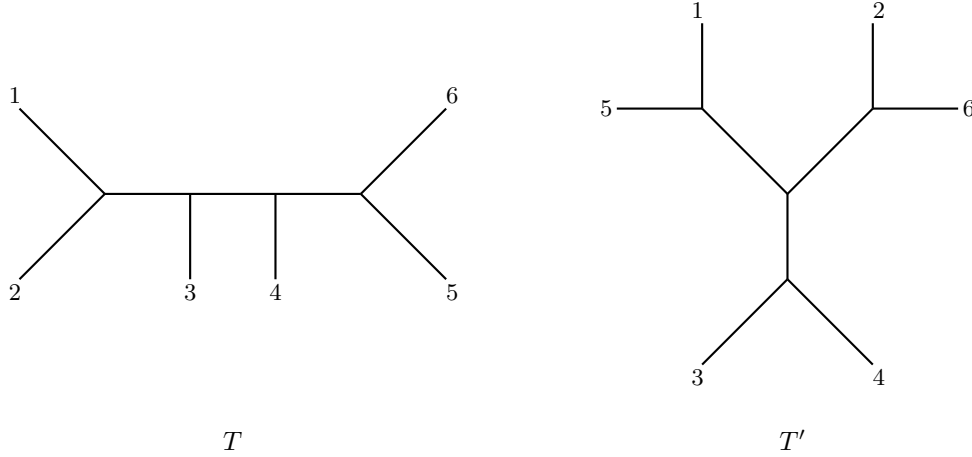


FIGURE 1. Two trees $T, T' \in UB(6)$. Note that T and T' are more than one TBR apart.

By calculating the parsimony score for the 32 different two-state characters (without loss of generality we set $f(1) := 0$) we can assign the values that Δ_1 can take and the probability of those values. $\Delta_1 = -2$ occurs precisely when $X_1(T) = 1$ and $X_1(T') = 3$ with probability $p = \frac{1}{32}$. $\Delta_1 = -1$ occurs precisely when $X_1(T) = 1$ and $X_1(T') = 2$ or $X_1(T) = 2$ and $X_1(T') = 3$ with probability $q = \frac{3}{32}$. $\Delta_1 = +1$ occurs precisely when $X_1(T) = 2$ and $X_1(T') = 1$ or $X_1(T) = 3$ and $X_1(T') = 2$

with probability $r = \frac{5}{32}$. Since $\Delta_1 = +2$ is not possible, $\Delta_1 = 0$ with probability $1 - (p + q + r) = \frac{23}{32}$. This leads to $\mathbb{P}(\Delta_1 < 0) = \mathbb{P}(\Delta_1 = -2) + \mathbb{P}(\Delta_1 = -1) = \frac{4}{32} < \frac{5}{32} = \mathbb{P}(\Delta_1 = +1) = \mathbb{P}(\Delta_1 > 0)$. Therefore $\mathbb{P}(X_k(T) < X_k(T')) < \mathbb{P}(X_k(T') < X_k(T))$ holds for $k = 1$ and the choice of T and T' shown in Fig. 1. In other words, the probability that the symmetric tree T is more parsimonious than the caterpillar tree T' (on a single random binary character) is higher than the probability that T' is more parsimonious than T .

□

3. MAXIMUM PARSIMONY TREES

Definition 3. [Maximum parsimony tree] Given a sequence $\mathcal{C} = (f_1, \dots, f_k)$ of characters on X , a phylogenetic tree T on X that minimizes $ps(\mathcal{C}, T)$ is said to be a *maximum parsimony (MP) tree* for \mathcal{C} . The corresponding ps-value is the *parsimony or MP score of \mathcal{C}* , denoted $ps(\mathcal{C})$.

Notation: Given $T \in UB(n)$, let $mp_k(T)$ denote the probability that T is an MP tree for $k \geq 1$ random two-state characters on $[n]$. That is

$$mp_k(T) := \mathbb{P}(X_k(T) \leq \min_{T' \in UB(n)} \{X_k(T')\}).$$

Notice that $mp_k(T)$ is not a probability distribution on $UB(n)$ since the positive probability of ties for the most parsimonious tree ensures that the $mp_k(T)$ values will sum to a value greater than 1.

Lemma 1. If $T_1, T_2 \in UB(n)$ have the same shape then $mp_k(T_1) = mp_k(T_2)$.

Proof. Let $k \geq 1$ and let f_1, \dots, f_k be two-state characters. Then $ps(f_1, \dots, f_k, T) = ps(f_1^\sigma, \dots, f_k^\sigma, T^\sigma)$, where σ is an element of the group S_n of permutations on the leaf set $[n]$ of T . Notice that the map $f = (f_1, \dots, f_k) \mapsto f^\sigma = (f_1^\sigma, \dots, f_k^\sigma)$ is a bijection, so the number of characters f for which T is an MP tree for f equals the number of characters f for which T^σ is an MP tree for f .

□

It follows from Proposition 1 and Lemma 1 that if $n \geq 3$ and $k = 1$, or if $k \geq 1$ and $n \leq 5$, then $mp_k(T)$ is constant for all $T \in UB(n)$. However, this does not hold more generally, as we now state.

Theorem 1. $mp_k(T)$ is not constant for all $T \in UB(n)$ when $n = 6$ and $k = 2$. In particular, any given caterpillar tree (like T in Fig. 1) has a higher probability of being an MP tree than a symmetric tree (like T' in Fig. 1).

The proof of Theorem 1 requires a detailed case analysis to identify the MP tree(s) for all pairs of characters (f_1, f_2) ; details are provided in the Appendix. The result is also confirmed by simulations, which are provided in the following section.

4. ASYMPTOTIC ANALYSIS

We first show that the bias exhibited in Proposition 2(ii) disappears asymptotically but the bias apparent in Theorem 1 does not.

Proposition 3. For all $T, T' \in UB(n)$ and all n :

$$\lim_{k \rightarrow \infty} \mathbb{P}(X_k(T) < X_k(T')) = \frac{1}{2}.$$

Proof. Let $T, T' \in UB(n)$ and $k \geq 1$, and let $\Delta_k = X_k(T) - X_k(T') = D_1 + D_2 + \dots + D_k$, where the random variable $D_i = ps(f_i, T) - ps(f_i, T')$ ($i = 1, \dots, k$) and the D_i are independent and identically distributed. Moreover $\mathbb{E}[D_i] = 0$ and D_i has a standard deviation σ that is strictly positive and finite. To see that $\sigma > 0$, note that $\sigma^2 \geq \mathbb{P}[D_i \neq 0]$ by Chebychev's inequality, and D_i is nonzero whenever f_i corresponds to a two-state character that has parsimony score 1 on one of the trees T, T' and parsimony score greater than 1 on the other tree (at least one such character must exist, since $T \neq T'$, and every tree is uniquely determined by its characters of parsimony score 1). We can now apply the standard Central Limit Theorem to deduce that for an asymptotically standard normal variable $Z_k = \frac{\Delta_k - \mathbb{E}[\Delta_k]}{\sigma \sqrt{k}}$,

we have:

$$\mathbb{P}(\Delta_k < 0) = \mathbb{P}\left(Z_k < \frac{0 - 0}{\sigma \cdot \sqrt{k}}\right) \xrightarrow{k \rightarrow \infty} \frac{1}{2}.$$

□

Finally, we consider the limiting behaviour of $mp_k(T)$ as $k \rightarrow \infty$, and present simulations that suggest that even for $n = 6$, this probability depends on the shape of the tree. It is easily shown that, for any $n > 1$, as $k \rightarrow \infty$, there is a unique most parsimonious tree, so $\sum_{T \in UB(n)} \lim_{k \rightarrow \infty} mp_k(T) = 1$ (see e.g. [17] (Theorem 4(2))). In other words, $\lim_{k \rightarrow \infty} mp_k(T)$ is a probability distribution on $UB(n)$. However, the additional claim there that $mp_k(T)$ is uniform on $UB(n)$ does not hold when $n = 6$ and when either $k = 2$ (Theorem 1) or, it seems, as $k \rightarrow \infty$, as we now explain.

4.1. Simulations. We used the computer algebra system *Mathematica* to generate alignments of lengths 2, 10, 100, 1,000, 10,000 and 100,000, respectively, by sampling characters for six taxa uniformly at random out of the 32 possible binary characters (we assume without loss of generality that the state of taxon 1 is fixed, say, to state 0, whereas all other taxa can choose states 0 or 1). For each alignment, we ran an exhaustive search through the tree space of 105 unrooted binary phylogenetic trees in order to find all MP trees. For each alignment length, we did 1,000 runs and we counted the average number of MP trees, as well as the number of times that each of the two tree shapes for six taxa (the caterpillar shape or the symmetric shape of T and T' in Fig. 1) were amongst the MP trees. We then calculated the ratio of the number of MP trees with a symmetric shape divided by the total number of MP trees. Note that this ratio should equal $\frac{1}{7} \approx 0.142857$ if both tree shapes were equally likely, because 15 out of the 105 possible trees have the symmetric shape. However, the last column of Table 1 reveals that only for the extremely short alignment of length 2 the ratio is close to this value in our simulations (and it is not exactly equal to it, by Theorem 1). Moreover, the ratio decreases away from $\frac{1}{7}$ as the alignment length increases (the small variation at alignment length 10,000 is within one standard deviation). This trend and the reported

values strongly suggest that the limiting value of $mp_k(T)$ is not uniform across all trees in $UB(6)$. Note also that column 2 of Table 1 is also consistent with the finding mentioned earlier that there will be a unique MP tree a with probability converging to 1 as k grows.

Al. length	av. # MP trees	# symmetric tree was MP	# caterpillar was MP	$\frac{\text{\# symmetric MP trees}}{\text{\# MP trees}}$
2	17.177	2,375	14,802	0.138266
10	3.908	365	3,543	0.0933982
100	1.622	119	1503	0.0733662
1,000	1.166	59	1107	0.0506003
10,000	1.053	57	996	0.0541311
100,000	1.013	46	967	0.0454097

TABLE 1. Overview of simulation results: For each alignment length, 1,000 runs were evaluated.

4.2. Concluding comments. In one sense, the two-state symmetric model is as favourable to all binary phylogenetic trees as is possible under maximum parsimony, since each tree has exactly the same probability distribution on the parsimony score of k random characters. Moreover, Proposition 3 shows that no one tree is any more likely to be an MP tree than another. It may seem somewhat surprising, therefore, that the distribution of MP trees is not uniform, even asymptotically; however this has a simple explanation. Although the characters are generated independently, and their parsimony scores is also independently distributed on any given binary tree, the MP binary tree is chosen once the k characters are given. Thus these characters are not independent random variables once we condition on a given tree being the MP tree for these characters. Moreover, once one moves away from the simple two-state model (for example, to the r -state symmetric model) even the uniformity of MP scores on fixed trees disappears [15]. In summary, while maximum parsimony on random data seems, in certain senses (described above), to favour each binary tree equally, the method nevertheless exhibits a bias towards trees with certain tree shapes.

4.3. Acknowledgments. We thank the Allan Wilson Centre for help funding this work. We also thank David Bryant for pointing out that MP trees might not be uniformly distributed on sequences of random characters.

REFERENCES

- [1] Aldous, D. (1995). Probability distributions on cladograms. In: Aldous, D., Pemantle, R.E. (Eds.), *Random Discrete Structures*. IMA Volumes in Mathematics and its Applications, vol. 76. Springer, pp. 1–18.
- [2] Aldous, D. (2001). Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Statistical Science* 16: 23–34.
- [3] Aldous, D., Krikun, M., Popovic, L. (2011). Five statistical questions about the tree of life. *Systematic Biology* 60: 318–328.
- [4] Bryant, D. (2004). The splits in the neighborhood of a tree. *Annals of Combinatorics* 8: 1–11.
- [5] Felsenstein, J. (2004). *Inferring Phylogenies*. Sinauer Associates.
- [6] Fitch, W. M. (1971). Towards defining the course of evolution: minimum change for a species tree topology. *Systematic Zoology*, 20: 406–416.
- [7] Foulds, L.R., and Graham, R. L. (1982). The Steiner problem in phylogeny is NP-complete. *Advances in Applied Mathematics*, 3: 43–49.
- [8] Harding, E.F. (1971). The probabilities of rooted tree shapes generated by random bifurcation. *Advances in Applied Mathematics* 3: 44–77.
- [9] Hartigan, J. A. (1973). Minimum mutation fits to a given tree. *Biometrics*, 29, 53–65.
- [10] Hey, J. (1992). Using phylogenetic trees to study speciation and extinction. *Evolution* 46, 627–640.
- [11] Holton, T.A., Wilkinson, M. and Pisani, D. (2014). The shape of modern tree reconstruction methods. *Systematic Biology*, 63(3):436–441.
- [12] Lambert, A, Stadler, T. (2013). Birth–death models and coalescent point processes: the shape and probability of reconstructed phylogenies. *Theoretical Population Biology* 90:113–128.
- [13] Semple, C. and Steel, M. (2003). *Phylogenetics*. Oxford University Press (Oxford Lecture Series in Mathematics and its Application).
- [14] Stadler, T. (2013). Recovering speciation and extinction dynamics based on phylogenies. *Journal of Evolutionary Biology*. 26:1203–1219.
- [15] Steel, M.A. (1993). Distributions on bicoloured binary trees arising from the principle of parsimony. *Discrete Applied Mathematics* 41(3): 245–261.
- [16] Vinh, L.S. Fuehrer, A. and von Haeseler, A. (2010). Random tree-puzzle leads to the Yule–Harding distribution. *Molecular Biology and Evolution* 28(2):873–877.
- [17] Zhu, S and Steel, M. (2013). Does random tree puzzle produce Yule–Harding trees in the many-taxon limit? *Mathematical Biosciences* 243: 109–116.

5. APPENDIX: PROOF OF THEOREM 1

We first recall some definitions and establish some preliminary lemmas.

Definition 4. [X -split, compatible]

An X -*split* is a bipartition of X into two nonempty subsets, written $A \mid B$. Given any phylogenetic X -tree T , if we delete any particular edge e of T and consider the leaf sets of the two connected components of the resulting disconnected graph we obtain an X -split, which is called a *split of T* (corresponding to e). If two X -splits $A \mid B$ and $A' \mid B'$ of the same unrooted phylogenetic X -tree have the property that one of the four intersections $A \cap A'$, $A \cap B'$, $B \cap A'$, $B \cap B'$ is empty, then $A \mid B$ and $A' \mid B'$ are said to be *compatible*. A two-state character f on X is said to be *compatible* with a phylogenetic X -tree T if $f^{-1}(0) \mid f^{-1}(1)$ is an X -split of T . Moreover, a pair of two-state characters f_1 and f_2 on X are said to be compatible with each other if f_1 and f_2 induce compatible X -splits (equivalently, if there exists a phylogenetic X -tree that f_1 and f_2 are compatible with). Finally, a two-state character f on X is *constant* if $f(x)$ takes the same value (0 or 1) for all $x \in X$.

The following result is easily established, with Part (b) following from the Split-Equivalence Theorem [13, Theorem 3.1.4].

Lemma 2.

- (a) A two-state character f is compatible with T if and only if $ps(f, T) = 1$.
- (b) A pair of two-state characters f_1 and f_2 are compatible if and only if there exists a tree $T \in UB(X)$ such that f_1 and f_2 are both compatible with T .

Lemma 3. For a phylogenetic X-tree T , and a pair $\mathcal{C} = (f_1, f_2)$ of two-state characters on X the following holds:

$$ps(\mathcal{C}, T) = \begin{cases} 0, & \text{if } f_1, f_2 \text{ are constant;} \\ 1, & \text{if } f_1 \text{ is compatible with } T \text{ and } f_2 \text{ is constant (or vice versa);} \\ 2, & \text{if } f_1, f_2 \text{ are compatible with } T; \\ \geq 3, & \text{otherwise.} \end{cases}$$

Proof. Let T be a tree with two constant two-state characters (f_1, f_2) . Then for both characters the parsimony score is 0 and therefore the maximum parsimony score is 0. Next, without loss of generality, let f_1 be compatible with T and f_2 constant, then

$$ps(\mathcal{C}, T) = \underbrace{ps(f_1, T)}_{=1, \text{ by Lemma 2(a)}} + \underbrace{ps(f_2, T)}_{=0, \text{ because } f_2 \text{ is constant}} = 1 + 0 = 1.$$

Now suppose that f_1 and f_2 are both not constant, but are compatible with T . By Lemma 2(a) the parsimony score of each character is 1, so that $ps(\mathcal{C}, T) = 2$. In all other cases we know that neither f_1 nor f_2 are constant, so $ps(f_1, T) \neq 0 \neq ps(f_2, T)$. Additionally, at least one of the two characters has a score of at least 2, because it is not compatible with T . Thereby the parsimony score is at least 3. \square

Lemma 4. For a pair $\mathcal{C} = (f_1, f_2)$ of two-state characters we have:

$$\min_{T \in UB(n)} \{ps(\mathcal{C}, T)\} = \begin{cases} 0, & \text{if } f_1 \text{ and } f_2 \text{ are constant;} \\ 1, & \text{if exactly one of } f_1 \text{ or } f_2 \text{ is constant;} \\ 2, & \text{if neither } f_1 \text{ nor } f_2 \text{ are constant, but } f_1 \text{ and } f_2 \text{ are compatible;} \\ 3, & \text{otherwise.} \end{cases}$$

Proof. If f_1 and f_2 are both constant then the MP score of this pair of characters on any tree is 0. Otherwise if exactly one of f_1 or f_2 is constant (without loss of generality, f_1 is

constant) then

$$\min_{T \in UB(n)} \{ps(\mathcal{C}, T)\} = \min_{T \in UB(n)} \{ps(f_1, T) + ps(f_2, T)\} = \min_{T \in UB(n)} \{0 + ps(f_2, T)\} = 1.$$

Now, suppose that neither f_1 nor f_2 are constant, but f_1 and f_2 are compatible with each other. From Lemma 2(b) there exists a tree $T \in UB(n)$ such that f_1 and f_2 are compatible with this tree, and so, Lemma 3 shows that $ps(\mathcal{C}, T) = 2$ and that T is an MP tree for \mathcal{C} . In the last case f_1 and f_2 are not constant and f_1 and f_2 are incompatible with each other. For the corresponding X -splits, $A_1 \mid B_1$ and $A_2 \mid B_2$, we may suppose that A_1 and A_2 correspond to 0, B_1 and B_2 correspond to 1. Consider any phylogenetic X -tree of the type shown in Fig. 2 with the following leaf sets (none of which is empty); $X = A_1 \cap A_2, Y = A_1 \cap B_2, W = B_1 \cap A_2, Z = B_1 \cap B_2$. Note that there is no tree with a lower score by Lemma 3.

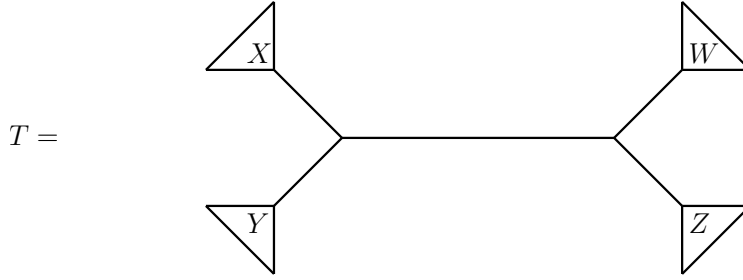


FIGURE 2. $T \in UB(n)$ with four disjoint subtrees X, Y, W and Z . The tree structure of X, Y, W and Z is unimportant.

For f_1 all leaves in X and Y are in state 0 and all leaves in W and Z are in state 1. For f_2 all leaves in X and W are in state 0 and all leaves in Y and Z are in state 1. Then the MP score of the two characters on this tree is $\min_{T \in UB(n)} \{ps(f_1, T) + ps(f_2, T)\} = 1 + 2 = 3$. \square

Proof of Theorem 1:

We describe an explicit counterexample for $n = 6, k = 2$, namely the trees T_1, T_2 in $UB(6)$ as shown in Fig. 3, for which we will show that $mp_2(T_1) > mp_2(T_2)$.

Let $F := \{f : X \rightarrow \{0, 1\}\}$ be the set of all two-state characters f on $X = \{1, \dots, 6\}$. Then $N := \{f : X \rightarrow \{0, 1\} : \# \text{ leaves in state 0 is either } 0, 1, n - 1 \text{ or } n\}$ is the set of all

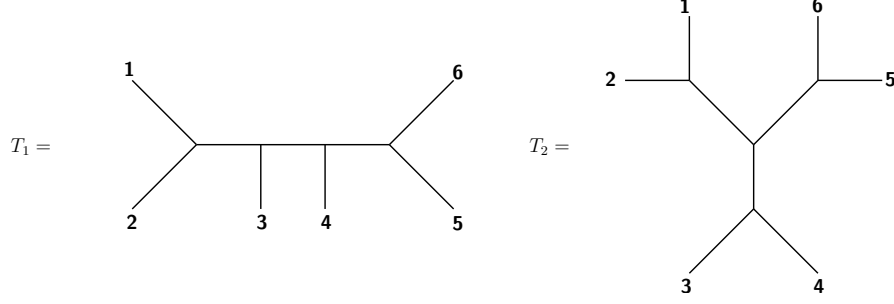


FIGURE 3. $T_1, T_2 \in UB(6)$ with different tree shapes.

non-informative two-state characters f on X . For each non-informative two-state character on a tree T , the character adds the same parsimony score to every tree (either 0 or 1). Furthermore, for any $T \in UB(n)$, define $I_j(T) := \{f \in F \setminus N : ps(f, T) = j\}$; $j = 1, 2, 3$. Thus, $I_j(T)$ is the set of all informative two-state characters f on X which have a parsimony score j on T , and when $n = 6$, F is the (disjoint) union of the four sets $I_1(T), I_2(T), I_3(T), N$. The number of characters in $I_1(T), I_2(T)$ and $I_3(T)$ is the same for any choice of $T \in UB(n)$ (this follows since the number of binary characters of parsimony score k is the same for each choice of $T \in UB(n)$ ([13], Theorem 5.6.2).

Now we have a look at all possible cases to choose f_1 and f_2 from $N, I_1(T), I_2(T)$ and $I_3(T)$. The following statements about various exclusive cases apply for any n , but we will specialise soon to $n = 6$ (since then the following seven cases exhaust every possibility).

Case 1: $f_1, f_2 \in N$. In this case, each tree $T \in UB(n)$ is an MP tree for this pair of characters, because there is no other tree with a lower score.

Case 2: $f_1 \in N$ and $f_2 \in I_1(T)$ or ($f_1 \in I_1(T)$ and $f_2 \in N$).

If the score of an informative character is 1, no tree achieves a better score than T , because only a non-informative character can have the score 0. Thus T is an MP tree in Case 2.

Case 3: $f_1 \in N$ and $f_2 \in I_2(T) \cup I_3(T)$ (or $f_1 \in I_2(T) \cup I_3(T), f_2 \in N$).

A non-informative f_1 contributes the same score to each tree $T \in UB(n)$. Moreover, when the score of an informative character is 2, this score can always be reduced. Therefore T with these characters is not an MP tree. Moreover, if f_1 has the score 1 and $f_2 \in I_3(T)$, the score of the T is 4, and so, by Lemma 4, T is not an MP tree. Finally, if f_1 has the score 0

and $f_2 \in I_3(T)$ one can always find a tree for f_2 which has a lower score. For this reason T is never an MP tree.

Case 4: $f_1, f_2 \in I_1(T)$.

As in Case 2 the scores of the characters cannot be improved by a tree different from T , so T is an MP tree.

Case 5: For $j = 2$ or $j = 3$, $f_1, f_2 \in I_j(T)$

A tree T with these characters has score 4 or score 6 and because of Lemma 4, T is never an MP tree.

Case 6: $f_1 \in I_3(T)$ and $f_2 \in I_1(T) \cup I_2(T)$ (or $f_1 \in I_1(T) \cup I_2(T)$ and $f_2 \in I_3(T)$).

A tree T with these characters has score 4 or score 5 and by Lemma 4, T is never an MP tree.

Case 7: $f_1 \in I_1(T)$ and $f_2 \in I_2(T)$ (or $f_1 \in I_2(T)$ and $f_2 \in I_1(T)$).

In this case, we need to further investigate whether or not T is an MP tree.

When $n = 6$ these represent all possible cases, and the only case where a different choice of $T \in UB(6)$ could affect whether or not T is an MP tree is Case 7, which we consider in detail now. To simplify the counting that follows, we may suppose, without loss of generality, that f_1 and f_2 both assign leaf 1 the state 0; moreover, for Case 7, we will just count the number of pairs of such characters (f_1, f_2) where $f_1 \in I_1(T)$ and $f_2 \in I_2(T)$ for $T \in UB(6)$.

The character f_1 can be described by making a change on a single edge α of T , while for f_2 we require two changes, on edges labelled β (we will see that these two β edges are not always uniquely determined by f_2). The placement of the two β edges in relation to the α edges falls into three scenarios, referred to as (a), (b) and (c) in Fig. 4 (circles in this figure denote leaves or subtrees).

Notice that in scenario (a) the splits induced by f_1 and f_2 are incompatible. Thus, since the MP score of T is 3, and this is best possible (by Lemma 4, since the splits are incompatible),

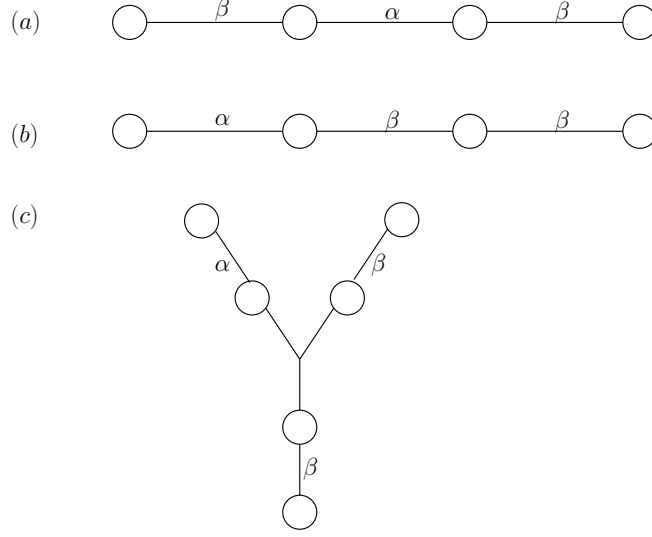


FIGURE 4. α : edge which corresponds to the single state change for $f_1 \in I_1(T)$ and β : two edges which correspond to the two state changes for $f_2 \in I_2(T)$.

so T is an MP tree under this scenario. In scenarios (b) and (c) the splits induced by f_1 and f_2 are compatible, and since the MP score of T of 3 is not best possible (again by Lemma 4, since the splits are compatible), T is not an MP tree. In summary, T is an MP tree if and only if scenario (a) applies. We thus want to count the number of pairs of characters (f_1, f_2) with $f_1 \in I_1(T)$ and $f_2 \in I_2(T)$ that correspond to scenario (a), and determine how this depends on the shape of the tree.

For T_1 there are three possible edges for α , so that $f_1 \in I_1(T)$ (see Fig. 5). To arrive at

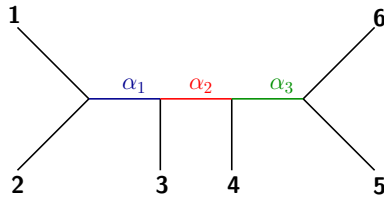


FIGURE 5. $\alpha_1, \alpha_2, \alpha_3$ are the three possible edges for T_1 , so that $f_1 \in I_1(T)$.

scenario (a), the two β edges must be on different sides of α . So we have $2 \cdot 6 = 12$ different options to place the β edges for each α_1 and α_3 . For α_2 we have $4 \cdot 4 = 16$ different options. But we are not only interested in how many places for changes we have. We rather want to count the possible two-state characters f_2 . So we have to check if we count some two-state

characters f_2 twice. We find that for every α_i ($i = 1, 2, 3$) we count exactly two characters twice (see Fig. 6). So for T_1 in Case 6 we get $2 \cdot 6 + 2 \cdot 6 + 4 \cdot 4 - 6 = 34$ pairs of two-state

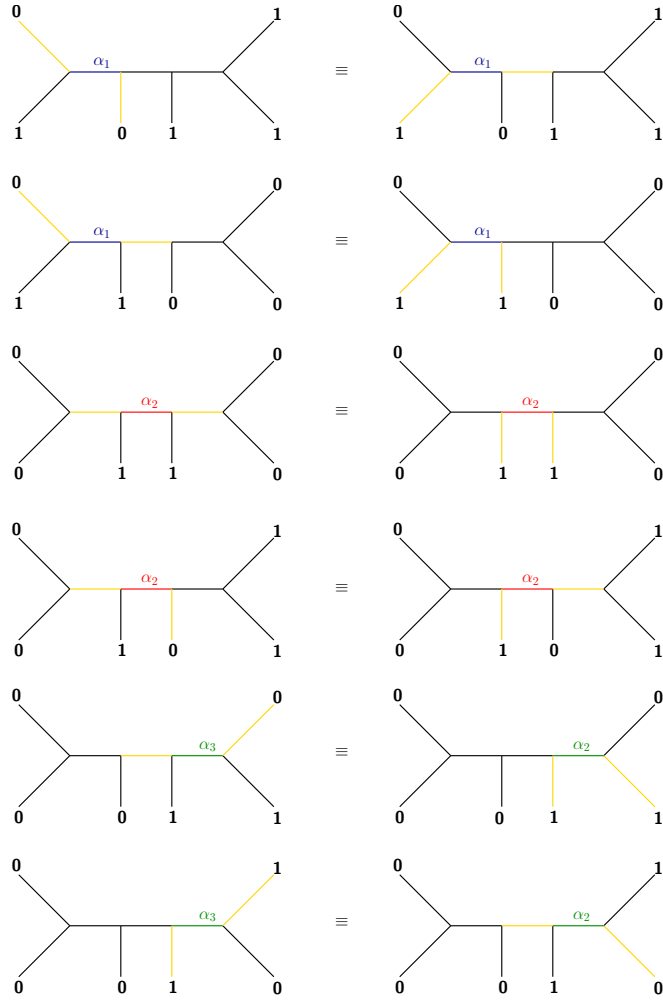


FIGURE 6. For every α_i ($i = 1, 2, 3$) we count two two-state characters twice.

characters f_1 and f_2 corresponding to scenario (a) (i.e. when T_1 is an MP tree).

Now we repeat this type of analysis for T_2 , where we can also find three possible edges for α ; $\alpha_1, \alpha_2, \alpha_3$ (see Fig. 7). But because of the symmetry of T_2 we just have to focus on one case. Here we focus on the edge α_1 ; the other two cases are analogous. To arrive at scenario (a) we have to place the β edges on different sides of the α edge. So we get $3 \cdot (2 \cdot 6) = 36$ ways to achieve this for a combination of one α and the two β edges. But once again we must check which two-state characters are counted twice. Here there are two cases for every possible α . The two cases for α_1 are shown in Fig. 8. Therefore we get $36 - 3 \cdot 2 = 30$ combinations of

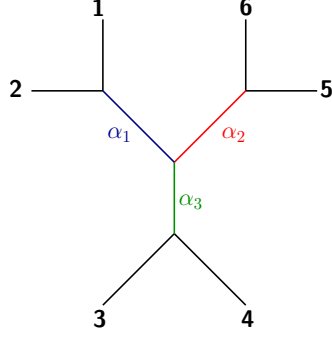


FIGURE 7. $\alpha_1, \alpha_2, \alpha_3$ are the three possible edges for T_2 , so that $f_1 \in I_1(T)$.

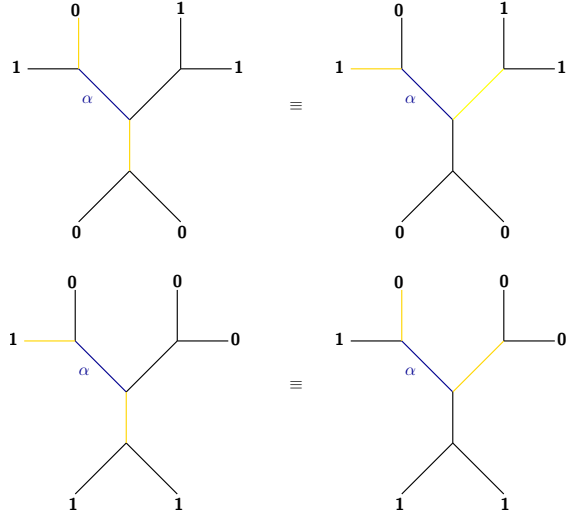


FIGURE 8. For every $\alpha_i (i = 1, 2, 3)$ we count two two-state characters twice. Here the two two-state characters we count twice for α_1 are shown.

f_1 and f_2 so that T_2 is an MP tree. Now we see that for T_1 there are more combinations of $f_1 \in I_1$ and $f_2 \in I_2$ to be an MP tree than for T_2 . For the reason that in all other cases the number of combinations of f_1 and f_2 to be an MP tree are the same, we can conclude that the probability that T_1 is an MP tree is higher than for T_2 . \square

ALLAN WILSON CENTRE, UNIVERSITY OF CANTERBURY, CHRISTCHURCH, NEW ZEALAND

DEPARTMENT FOR MATHEMATICS AND COMPUTER SCIENCE, ERNST-MORITZ-ARNDT UNIVERSITY,
GREIFSWALD, GERMANY